DOCUMENT RESUME

ED 159 599 CS 004 266

AUTHOR Petrosky, Anthony R.
TITLE The 3rd National Assessment of Reading and Literature Versus Norm- and Criterion-Referenced Testing.
PUB DATE May 78
NOTE 13p.; Paper presented at the Annual Meeting of the International Reading Association (23rd, Houston, Texas, May 1-5, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; Educational Assessment; Elementary Secondary Education; *Evaluation Methods; Literature; *National Competency Tests; National Norms; *Norm Referenced Tests; Reading Tests; Standardized Tests; Testing Programs; *Test Interpretation
IDENTIFIERS *Domain Referenced Tests; *National Assessment of Educational Progress

ABSTRACT

In discussing the third national assessment of reading and literature, four major points can be made. First, norm-referenced tests and criterion-referenced tests ignore serious ethical and measurement problems, namely, we don't know enough about individual differences to do such testing and the outcome, social class tracking, is ethically repulsive. Second, comprehending and interpreting literary texts is a subset of reading not separate from it. Third, descriptive information from the national assessment survey is useful in considering notions about developmental differences in the ways students interpret and evaluate which are attributable to developmental growth, schooling, and personal inclinations. Fourth, domain-referenced assessment is the best standardized procedure for finding out what students know and what teachers and schools can do. It eliminates the ambiguity created by behavior factors by representative sampling from a well-defined set of tasks, by referring to the logical relationship between a set of items in a test and a well-defined domain represented by those items, and by estimating the kinds of behavior students are capable of within a defined domain. (TJ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED159599

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Anthony R. Petrosky

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

THE 3rd NATIONAL ASSESSMENT OF READING AND LITERATURE VERSUS NORM-AND CRITERION-REFERENCED TESTING

Anthony R. Petrosky
University of Pittsburgh
The National Assessment of Educational Progress

Introduction

I want to make 4 major points in my paper: (1) proponents of national norm- and criterion-references testing--those people with the seemingly easy answers to guaranteeing competency in basic skills--systematically ignore serious ethical and measurement problems peculiar to assessing mental processes like reading and literary experience, (2) the ability to comprehend and interpret literary texts--fiction and non-fiction--is a subset of reading, (3) descriptive information from national assessment surveys is useful for establishing or rejecting notions about developmental differences in learning and, consequently, as the argument goes, in teaching, and (4) domain-referenced assessment, not norm- or criterion-referenced, is the best standardized procedure for finding out what students know and can do and what teachers and schools do.

CS804266

General Problems with Testing

For my first point, I want to discuss what was one of the largest systematic testing situations in the English speaking world in order to illustrate a very real ethical problem that permeates this kind of assessment. If you were an adolescent growing up in Great Britain, chances are that you would have taken the 11+ Exam along with the rest of the kids in the country to see what kind of advanced schooling you were cut out for. Some of your friends, of course, would have been worth an investment and could, then, proceed along the academic path to professional careers. Other less fortunates would have stayed in the lower ranks training for the less complicated and demanding labors best suited to them. Given the two most powerful predictors of academic achievement--SES and schooling--three quite obvious conditions could have dealt you a losing hand: (1) you might have been born into a low SES rank, (2) you could have lacked the abilities needed to advance intellectually, and (3) you could have done poorly on the exam. Certainly, these are not all the possibilities an individual confronted as he progressed through his British schooling, but these 3 illustrations did account for a great deal of academic failure in Great Britain according to David Holbrook (1964). Generally, people born into low SES ranks stay there, and people who are not very smart do not progress. These two phenomena traditionally resist treatments, at least that is the case on this tiny planet. The third phenomenon, doing poorly on the exam, could be the result of many factors--some profound, some trivial. Of course, someone from a low SES rank could have done well on the 11+ and advanced but, again, according to Holbrook this was not generally the case. It was not the case, because, as you may suspect,

in order to do well on this kind of exam, you need the resources and experiences that are not available to lower class people. Obviously this isn't a new or profound observation. This old dog of an argument has been chasing its tail for some time. But, the salient point is that there are two powerful ways to reap the benefits of our industrialized world. One is birth, the other is schooling. If you aren't born into the right social class, your future hinges, for the most part, on schooling (of course there's always luck). Good and influential schooling in Great Britain was parceled out by normative testing--the 11+ Exam. How you scored determined the kind of schooling you got.

I suppose there is an argument somewhere that this kind of testing can benefit adolescents by giving them a chance to prove their worth. But that argument relies on the assumption that norm- and criterion-referenced tests are good, in the sense that they are valid, unbiased, and reliable. Tierney (1978) argues that they are not, and I agree. We know very little about measurement and individual differences. We certainly do not know enough to carry off a national norm- or criterion-referenced assessment of reading with intellectual and social integrity. There is no question as to whether national testing can be a way of social class tracking. The question is do we want to do it? I hope not. If we insist on sealing our children's careers early in their lives with a test like Britain's old 11+ Exam, then we must consider the heart of a very old and perplexing theoretical problem: can we assume that what we know is always expressible and therefore measurable? Socrates said it best, I think, when he asked Laches in the dialogue by the same name, to define courage. Although he is a man of courage himself

and has seen courage many times in others, Laches cannot do it, and neither can Socrates. Plato makes his point with a twist of irony when he proposes early on in Laches, the instigating statement: "And that which we know we must surely be able to tell." Clearly, this is not always the case. The fundamental issue in the dialogue is that Laches assumes like most assessment assumes, that conscious articulation of unconscious knowledge and processes, like acting with courage--or reading--, is possible and, then, accurate. The truth is that our methods and measures do not usually match our purposes; it is difficult to construct measures that do what we want them to, and yet we continually try to assess things we know very little about. If we look at published tests of reading and literature--an opportunity I availed myself of this past year--we see that measurement usually takes the form of testing component skills as if they could all be added up to form a sum total of someone's ability. This kind of atomistic thinking about integrated and complex mental processes takes the cash in the market place.

Let me reiterate my two main points. We do not know enough about individual differences, mental processes, and measurement to do national norm- or criterion-referenced testing of reading, and the final outcome of these kinds of testing--social class tracking--is ethically repulsive. Second, research is beginning to confirm what both common sense and intellectual integrity might lead us to suspect, which is that self-consciousness and memory recall--the two main attributes of reading and literature tests--are not necessary or sufficient conditions for the reading process.

Literary Experience

I have always been puzzled by the separation of reading and literature. Even more astonishing is the assertion that reading is a process and literature a content. Why do we continually reduce ad absurdum to the point where we talk about processes without content or vice versa? It seems to me that the process of reading must always have a content and as long as we are reading print--be it fiction or non-fiction--the process is a matter of what the brain does with the marks on the page. If Frank Smith (1971) is correct, and I think the evidence he presents is very convincing, then it is very reasonable to say that the brain processes all print for meaning in the same way. The differences between fiction and non-fiction, if there are any, must surely be the result of what we do after we process the print for meaning. On the other hand, there is an argument, I think, for asserting that readers do different things in response to fiction than they do in response to most non-fiction (straight information). Fiction, according to Norman Holland (1973, 1975) opens the way for readers to indulge in personal interpretations and evaluations which often lead to a dialectic--a sharing of thoughts and feelings that can, quite easily, lead to practice in decision making and consensus oriented conventions. And there is some literary reading that requires special learning on the part of the reader, especially where literary devices and structures like metaphor and hyperbole are involved. But, from what we know about literary devices and the fluent processing of print (Pollio et al, 1977), it is reasonable to assert that perceiving literary devices in the fluent processing of print is no more a conscious, analytic skill than perceiving topic sentences in the reading

of expository prose. The devices are there; we learn to depend on them, but we do not interrupt the reading process to perceive them as entities in themselves. Ordinarily, these devices are responded to in an unmediated way, subconsciously or automatically in the reading process, so that readers might successfully infer a meaning for a particular text but not recognize what particular devices or structures contributed to their responses and in what ways. Literary devices and structures should not, then, be assessed by testing for the vocabulary of literary criticism (e.g., what is alliteration?), since readers can depend on the phenomenun at a sophisticated level of understanding without knowing the official label. It is for these reasons that we incorporated literary devices and structures into the comprehension objective of the 3rd national assessment of reading and literature. We think, in addition, that literary devices and structures--that is, the comprehension of them--can be assessed in lexical, propositional, and textual (whole text) contexts and the comprehension objective of the assessment reflects this notion by parceling comprehension into these three basic categories. We want to know how students read all types of texts and for assessment purposes, the distinction between reading and literature is a fruitless one. Comprehension--be it of literary texts or not--is relative to the reader, the text, and the task at hand and the ability to comprehend literary texts is a subset of reading.

Still, there are some specific literary type activities we want to know about. Although these activities, interpretation and evaluation, are not solely literary in nature, they are usually encountered in the literature classroom which is, in all probability, unfortunate for the students who

take reading and not literature. In any case, interpretation and evaluation, as we define them, require readers to imagine the significance of passages in relation both to the entire text and to themselves. Interpretation and evaluation are marked by such activities as conscious manipulation of the text and the readers' expressed responses. Interpretation refers to the explanation or elucidation of meaning in the text, often in highly personal ways, after the processing of print for meaning occurs; and evaluation refers to judgment. The assessment of these activities is designed to show whether readers interpret and evaluate texts according to author's intentions, aesthetic criterion, personal experiences, or other factors. The crux of the assessment in this domain is descriptive in nature. We want to collect information that might shed some light on developmental differences and preferences in interpretation and evaluation. Certainly, there is no better representative national sample of 9, 13, and 17 year olds than NAEP's for building this kind of descriptive profile. And this is my third major point: we need to know if there are differences in the ways students interpret and evaluate that are attributable to developmental growth, schooling, and personal inclinations. Unlike the IEA study of Literature Education in Ten Countries (Purves, 1975), we are not primarily focusing on the influence of schools; we want to know what students know and can do, and, it seems to me, what teachers and schools do is part and parcel of what students know and do. The question is whether we can attribute specific behaviors to developmental differences as opposed to schooling differences. I have my doubts about making this distinction, but at least we can collect information to construct accurate descriptive profiles.

To this end we have identified three formats to measure interpretive and evaluative responses. The first approach asks readers to select statements that are most similar to their interpretative and evaluative responses after reading a selection. The statements represent retellings, generalizations, personalizations, symbolic responses, and misunderstandings. We are anticipating that this format will yield patterns of preferences specific to the students' age groups. We are in effect hypothesizing that 9, 13, and 17 year olds will show distinct preference patterns by age.

The second format takes advantage of other questions in the assessment. After a question that asks readers to infer something like tone or character, we ask them to explain that inference. Unlike the first format where students are limited to selecting responses from a limited set; the second format is completely open-ended and can be scored using the same scheme as the first format or by categorizing the responses as primarily text related (basing explanations on textual material), reader related (basing explanation on personal experiences), or unrelated. By allowing ourselves the flexibility of these scoring guides we can determine whether students prefer response statements that are retellings, generalizations, personalizations, symbolic renderings, and whether the explanation they give is related more to the text or to their personal experiences.

A third approach asks readers to list some elements of a good story, a good editorial, a good advertisement, or any other genre. After completing the list, they read a passage and are asked to evaluate it. We then ask them to list reasons why they rated it as they did. Calling not only for an evaluation but also for an explanation of the evaluative process,

this format should yield information on how well readers understand and articulate the criteria they use for evaluation.

Domain-referenced Assessment

The procedures for constructing test specifications for the Year 3 Reading and Literature Assessment are adapted from the work of Popham (1976), Engel (1977), and Martuza (1977) on domain-referenced testing. The entire procedure, including objective amplification--a process where test specifiers systematically eliminate ambiguity in directions to item writers--is meant to facilitate item writing. Besides the description of the activities to be assessed, rules and guidelines for constructing test items are spelled out for all of the controlling dimensions of the domain (e.g. response description, item format, scoring criteria, etc.) to create congruent test items within the specifications.

Domain-referenced assessment differs from norm- and criterion-referenced assessment in three important ways: (1) the ambiguity of content by behavior matrices usually used for constructing norm- and criterion-referenced tests is eliminated by representative sampling of tasks from a well-defined set of tasks, (2) whereas criterion-referenced refers to the way the examinee's test score is interpreted in terms of preset performance standards, domain-referenced refers to "the logical relationship which exists between a set of items in a test and a well-defined domain represented by those items" (Martuza, 1977), and (3) domain-referenced tests can be used to estimate the kinds of behaviors students are capable of within the defined domain.

Since NAEP does not report individual scores and reports findings by age, geographic region, sex, race, etc., it is probably more appropriate to refer to the assessment as a domain-referenced survey rather than a test. It is my argument that we can best find out what students know and can do by using this type of domain-referenced survey to describe the kinds of activities that students are capable of within a well-defined domain like Interpretation and Evaluation. The power of a domain-referenced assessment rests in the carefulness necessary to define the domain and amplify the objectives. The power of the NAEP Year 3 Reading and Literature Assessment rests not only in its use of a domain-referenced approach to objective amplification and item writing but also, and perhaps more importantly, in its integration of reading and literature throughout the three major objectives: valuing, comprehension, and interpretation. The descriptive nature of a good portion of the data that the assessment will yield can only add to the credibility of NAEP exercises and the overall Reading and Literature Assessment. Writers who interpret and report the results will be able to speak in specific and clear ways about the findings for individual items and clusters of items.

Summary

It is heartening in these times of competency based everything and accountability to see NAEP perfecting its survey instruments and carefully avoiding the serious problems of norm- and criterion-referenced tests. From a theoretical perspective, the integration of reading and literature for assessment purposes is a strong and important step forward, away from atomistic notions about human understanding and human mental abilities.

Throughout my paper I have painted a picture of NAEP assessment procedures as being much more desirable than the alternatives--norm- and criterion-referenced testing. I have spent a good deal of time discussing the assessment of interpretation and evaluation because I think this is the kind of information we, as a nation, need to know in order to make confident decisions about such things as developmental differences, schooling differences, and, finally, the effects of these on curriculum building.

Perhaps the most important dimension to NAEP is the example we set by the kind of assessment we put together. Theoretical frameworks and carefulness in constructing test specifications and test items speak well of a project that must cater to all sorts of political whims. The 3rd national assessment of reading and literature proceeds far beyond its predecessors in both the construction of test specifications and item writing.

References

Engel, J.D. Domain specifications. Working paper. Chicago: University of Illinois, 1977.

Holbrook, D. English for the rejected: Training literacy in the lower streams of the secondary school. Cambridge University Press, 1964.

Holland, N.N. Poems in persons: An introduction to the psychoanalysis of literature. New York: Norton, 1973.

Holland, N.N. Five readers reading. New Haven: Yale University Press, 1975.

Martuza, V.R. Applying norm-referenced and criterion-referenced measurement in education. New York: Allyn and Bacon, 1977.

Popham, W.J. Preparing criterion-referenced test specifications. Paper presented at AERA, 1976.

Pollio, H., Barlow, J., Fine, H., and Pollio, M. Psychology and the poetics of growth: Figurative language in psychology, psychotherapy, and education. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1977.

Purves, A.C. Literature education in ten countries. New York: John Wiley, 1975.

Smith, F. Understanding reading: A psycholinguistic analysis of reading and learning to read. New York: Holt, Rhinehart, and Winston, 1971.

Tierney, R. Assessing performance in reading. Paper presented at IRA, 1978.